(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
9 August 2001 (09.08.2001)

PCT

(10) International Publication Number
WO 01/57248 A2

(51) International Patent Classification[7]: C12Q 1/68

(21) International Application Number: PCT/GB01/00407

(22) International Filing Date: 31 January 2001 (31.01.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0002310.1 1 February 2000 (01.02.2000) GB

(71) Applicant: SOLEXA LTD. [GB/GB]; Chesterford Research Park, Little Chesterford, Saffron Walden, Essex CB10 1XL (GB).

(72) Inventors: BALASUBRAMANISN, Shankar; University of Cambridge, Department of Chemistry, Lensfield Road, Cambridge CB2 1EW (GB). BENTLEY, David; Sanger Centre, The Wellcome Trust Genome, Campus, Hinxton, Cambridge CB10 1SA (GB).

(74) Agent: GILL JENNINGS & EVERY; Broadgate House, 7 Eldon Street, London EC2M 7LH (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/57248 A2

(54) Title: POLYNUCLEOTIDE ARRAYS AND THEIR USE IN SEQUENCING

(57) Abstract: A device for use in polynucleotide sequencing procedures comprises an array of polynucleotide molecules capable of interrogation and immobilised on a solid surface, wherein the array has a density which allows the molecules to be individually resolved. The molecules on the array comprise a target polynucleotide and a primer sequence, the primer being maintained in spatial relationship to the target, both being linked via a covalent bond. The device may be used in genotyping experiments, in which an array of genomic DNA is prepared to allow multiple sequencing reactions to be performed.

POLYNUCLEOTIDE ARRAYS AND THEIR USE IN SEQUENCING

Field of the Invention

This invention relates to devices having an array of polynucleotides, and to the sequencing of polynucleotides. In particular, this invention discloses methods for determining the sequence of arrayed polynucleotides.

Background to the Invention

Advances in the study of molecules have been led, in part, by improvement in technologies used to characterise the molecules or their biological reactions. In particular, the study of the nucleic acids DNA and RNA has benefitted from developing technologies used for sequence analysis and the study of hybridisation events.

An example of the technologies that have improved the study of nucleic acids, is the development of fabricated arrays of immobilised nucleic acids. These arrays typically consist of a high-density matrix of polynucleotides immobilised onto a solid support material. Fodor *et al*, Trends in Biotechnology (1994) 12:19-26, describes ways of assembling the nucleic acids using a chemically sensitized glass surface protected by a mask, but exposed at defined areas to allow attachment of suitably modified nucleotide phosphoramidites. Fabricated arrays may also be manufactured by the technique of "spotting" known polynucleotides onto a solid support at predetermined positions (e.g. Stimpson *et al* PNAS (1995) 92:6379-6383).

A further development in array technology is the attachment of the polynucleotides to the solid support material to form single molecule arrays. Arrays of this type are disclosed in WO-A-00/06770. The advantage of these arrays is that reactions can be monitored at the single molecule level and information on large numbers of single molecules can be collated from a single reaction.

Although these arrays offer particular advantages in sequencing experiments, the preparation of arrays at the single molecule level may be more difficult than at the multi-molecule level, where losses of target polynucleotide can be tolerated due to the multiplicity of the array. There is therefore a need

for improvements in the preparation of the single molecule arrays for sequencing procedures.

Furthermore, for DNA arrays to be useful their sequences must be determined. US-A- 5302509 discloses a method to sequence polynucleotides immobilised on a solid support. The method relies on the incorporation of fluorescently-labeled, 3'-blocked bases A, G, C and T to the immobilised polynucleotide, in the presence of DNA polymerase. The polymerase incorporates a base complementary to the target polynucleotide, but is prevented from further addition by the 3'-blocking group. The label of the incorporated base can then be determined and the blocking group removed by chemical cleavage to allow further polymerisation to occur.

Sequencing polynucleotides on a solid support can be difficult as the polynucleotide is typically bound to the solid support via a hybridisation reaction with a support-bound complement. Occasionally the conditions used in the sequencing protocol result in disruption to the bonds formed on hybridisation and the target polynucleotide is removed from the array.

Canard *et al.*, Gene (1994) 148:1-6, discloses a DNA sequencing procedure using a relatively harsh alkali denaturation step to remove fluorescent labels from nucleotides incorporated onto a template. The alkali treatment results in melting of the primer-template duplex, and Canard suggests providing a hairpin-like primer to the template to overcome this problem. The hairpin-like primer is suggested only in the context of the harsh alkali treatment with a homogeneous sample of template DNA.

WO-A-98/20019 discloses compositions and methods for the preparation of nucleic acid arrays. The general disclosure relates to the preparation of high density multi-molecule arrays, achieved by immobilising polynucleotides on microscopic beads attached to a solid support. Many different uses are proposed for the arrays.

WO-A-97/08183 relates to nucleic acid capture molecules. Hairpin loop structures are disclosed, but only in the context of capture molecules for use in hybridisation-based nucleic acid detection methods.

WO-A-97/04131 discloses primer amplification experiments carried out with polynucleotide hairpin structures. The general teaching is of methods for amplification of the hairpin structures, and it is clear that the structures contain a polynucleotide sequence to which a separate primer sequence is hybridised. The primer associates with the hairpin only through the hybridisation reaction.

Summary of the Invention

The present invention is based on the realisation that sample preparation and sequence analysis procedures can be improved if the target polynucleotide is maintained in spatial relationship to the primer, and the target and primer are linked via a bond stronger than that of hybridisation.

According to a first aspect of the invention, a device comprises an array of polynucleotide molecules immobilised on a solid support, wherein each molecule comprises a polynucleotide duplex linked via a covalent bond to form a hairpin loop structure, one end of which comprises a target polynucleotide, and the array has a surface density which allows the target polynucleotides to be individually resolved.

According to a further aspect of the invention, the device of the invention is used in a procedure to determine the sequence of the target polynucleotide.

According to a third aspect, a method for the preparation of the device comprises ligating a target polynucleotide to the 5'-end of a first molecule capable of forming a duplex as defined above, and immobilising the first molecule to the solid surface either before or after ligation.

The present invention provides a method for capturing target polynucleotides from solution and immobilising them on an array to form a stable primer-template complex. The target polynucleotide is maintained on the array together with the primer by bonds stronger than those formed in hybridisation, and so conditions which would otherwise result in the target or primer being dissociated from the array, may be used. This may be of particular use in genotyping experiments, where an array of genomic DNA is prepared, to allow multiple sequencing reactions to be performed.

Description of the Invention

The present invention is the formation of arrays of polynucleotides, the polynucleotides comprising a hairpin loop structure, one end of which comprises a target polynucleotide, the other end comprising a relatively short polynucleotide capable of acting as a primer in the polymerase reaction.

The arrays are typically high density arrays, and there are preferably at least 10^3 molecules/cm^2, more preferably at least 10^5 molecules/cm^2 and most preferably 10^6-10^9 molecules/cm^2. The molecules immobilised on the solid support surface are also preferably at a density so that they can be considered to be single molecules, i.e. each can be individually resolved. The term "individually resolved" is used herein to specify that, when visualised, it is possible to distinguish one molecule on the array from its neighbouring molecules. Separation between individual molecules on the array will be determined, in part, by the particular technique used to resolve the individual molecules. It will usually be the target polynucleotide portion that is individually resolved, as it is this which will be interrogated, e.g. by the incorporation of detectable bases.

Apparatus used to image arrays are known to those skilled in the art. For example, a confocal scanning microscope may be used to scan the surface of the array with a laser, to image directly a fluorophore incorporated on the individual arrayed molecule by fluorescence. Alternatively, a sensitive 2-D detector, such as a change-coupled detector, can be used to provide a 2-D image representing the individual molecules on the array. Resolving single molecules on the array with a 2-D detector is possible if adjacent molecules are separated by a distance of approximately at least 250nm, preferably at least 300nm and more preferably by at least 350nm.

The arrayed polynucleotide molecules each comprise a polynucleotide duplex which is used to retain a primer and a target polynucleotide in spatial relationship to each other. This ensures that the primer is able to perform its priming function during a polymerase-based sequencing procedure, and is not removed during any washing step in the procedure. The target polynucleotide

is capable of being interrogated. The term "capable of interrogation" refers to the ability of the target to act as a template during the polymerase reaction. The target polynucleotide is therefore interrogated by the incorporation of bases and the polymerase enzyme.

The duplex forms a hairpin loop structure, and the primer and target polynucleotides are comprised at the respective ends of the hairpin loop structure. The term "hairpin loop structure" refers to a molecular stem and loop formed from the hybridisation of complementary polynucleotides that are covalently linked at one end. The stem comprises the hybridised polynucleotides and the loop is the region that links the two complementary polynucleotides. Anything from a 10 to 20 (or more) base pair double-stranded (duplex) region may be used to form the stem. In one embodiment, the structure may be formed from a single-stranded polynucleotide having complementary regions. The loop in this embodiment may be anything from 2 or more non-hybridised nucleotides. In a second embodiment, the structure is formed from two separate polynucleotides with complementary regions, the two polynucleotides being linked (and the loop being formed) by a linker moiety. The linker moiety forms a covalent attachment between the ends of the two polynucleotides. Linker moieties suitable for use in this embodiment will be apparent to the skilled person. For example, the linker moiety may be polyethylene glycol (PEG).

The term 'polynucleotide' is intended to refer to nucleic acids in general, including DNA, RNA and synthetic analogs, e.g. PNA. DNA is preferred.

Preparation of the devices to form the target polynucleotide on the array may be carried out by methods known to those skilled in the art. The target polynucleotide will usually be one to which a sequence determination is to be made.

There are many different ways of forming the hairpin structure to incorporate the target polynucleotide. However, a preferred method is to form a first molecule capable of forming a hairpin structure, and ligate the target polynucleotide to this. Ligation may be carried out either prior to or after immobilisation to the solid support. The resulting structure comprises the single-

stranded target polynucleotide at one end of the hairpin and a primer polynucleotide at the other end.

In one embodiment, the target polynucleotide is genomic DNA purified using conventional methods. The genomic DNA may be PCR-amplified or used directly to generate fragments of DNA using either restriction endonucleases, other suitable enzymes, a mechanical form of fragmentation or a non-enzymatic chemical fragmentation method. In the case of fragments generated by restriction endonucleases, hairpin structures bearing a complementary restriction site at the end of the first hairpin may be used, and selective ligation of one strand of the DNA sample fragments may be achieved by one of two methods.

Method 1 uses a first hairpin whose restriction site contains a phosphorylated 5' end. Using this method, it may be necessary to first de-phosphorylate the restriction-cleaved genomic or other DNA fragments prior to ligation such that only one sample strand is covalently ligated to the hairpin.

Method 2: in the design of the hairpin, a single (or more) base gap can be incorporated at the 3' end (the receded strand) such that upon ligation of the DNA fragments only one strand is covalently joined to the hairpin. The base gap can be formed by hybridising a further separate polynucleotide to the 5'-end of the first hairpin structure. On ligation, the DNA fragment has one strand joined to the 5'-end of the first hairpin, and the other strand joined to the 3'-end of the further polynucleotide. The further polynucleotide (and the other strand of the fragment) may then be removed by disrupting hybridisation.

In either case, the net result should be covalent ligation of only one strand of a DNA fragment of genomic or other DNA, to the hairpin. Such ligation reactions may be carried out in solution at optimised concentrations based on conventional ligation chemistry, for example, carried out by DNA ligases or non-enzymatic chemical ligation. Should the fragmented DNA be generated by random shearing of genomic DNA or polymerase, then the ends can be filled in with Klenow fragment to generate blunt-ended fragments which may be blunt-end-ligated onto blunt-ended hairpins. Alternatively, the blunt-ended DNA fragments may be ligated to oligonucleotide adapters which are designed to

allow compatible ligation with the sticky-end hairpins, in the manner described previously.

The hairpin-ligated DNA constructs may then be covalently attached to the surface of a solid support to generate a single molecule array (SMA), or ligation may follow attachment to form the array.

The arrays may then be used in procedures to determine the sequence of the target polynucleotide. In the case that the target fragments are generated via restriction digest of genomic DNA, the recognition sequence of the restriction or other nuclease enzyme will provide 4, 6, 8 bases or more of known sequence (dependent on the enzyme). Further sequencing of between 10 and 20 bases on the SMA should provide sufficient overall sequence information to place that stretch of DNA into unique context with a total human genome sequence, thus enabling the sequence information to be used for genotyping and more specifically single nucleotide polymorphism (SNP) scoring.

Simple calculations have suggested the following based on sequencing a 10^7 molecule SMA prepared from hairpin ligation: For an 8 base pair recognition sequence, a single restriction enzyme will generate approximately 10^6 ends of DNA. If a stretch of 13 bases is sequenced on the SMA (i.e. 13×10^6 bases), approximately 13,000 SNPs will be detected. One application of such a sample preparation and sequencing format would in general be for SNP discovery in pharmaco-genetic analysis. The approach is therefore suitable for forensic analysis or any other system which requires unambiguous identification of individuals to a level as low 10^3 SNPs.

It is of course possible to sequence the complete target polynucleotide.

The sequencing method that is used to characterise the bound target may be any known in the art that measures the sequential incorporation of bases onto an extending strand. A suitable technique is disclosed in US-A-5302509 requiring the monitoring of sequential incorporation of fluorescently-labeled bases onto a complement using the polymerase reaction. Alternatives will be apparent to the skilled person. Suitable reagents, including fluorescently-labeled nucleotides will be apparent to the skilled person.

Immobilisation of the polynucleotides to the solid support may be carried out by any method known in the art, provided that covalent attachment is achieved. The solid support may be any of the conventional supports used in "DNA chips", e.g. glass slides, ceramic, silicon or plastics materials. The support will typically be a flat planar surface, usually of approximately $1 cm^2$.

The single molecule array can be prepared by contacting a suitably prepared solid support with a dilute solution containing the polynucleotides to be arrayed. Techniques for preparing the arrays are detailed in WO-A-00/06770.

<u>Example</u>

This Example illustrates the preparation of single molecule arrays by direct covalent attachment of hairpin loop structures to glass.

A solution of 1% glycidoxypropyltrimethoxy-silane in 95% ethanol/5% water with 2 drops H_2SO_4 per 500 ml was stirred for 5 minutes at room temperature. Clean, dry Spectrosil-2000 slides (TSL, UK) were placed in the solution and the stirring stopped. After 1 hour the slides were removed, rinsed with ethanol, dried under N_2 and oven-cured for 30 min. at 100°C. These 'epoxide' modified slides were then treated with 1 µM of labelled DNA (5'-Cy3-CTGCTGAAGCGTCGGCAGGT-heg-aminodT-heg-ACCTGCCGACGCT-3') (SEQ ID NOS. 1 and 2) in 50 mM potassium phosphate buffer, pH 7.4 for 18 hours at room temperature and, prior to analysis, flushed with 50 mM potassium phosphate, 1 mM EDTA, pH 7.4. The coupling reactions were performed in sealed teflon blocks under a pre-mounted coverslip to prevent evaporation of the sample and allow direct imaging.

The DNA structure was designed as a self-priming template system with an internal amino group attached as an amino deoxy-thymidine held by two 18 atom hexaethylene glycol (heg) spacers, and was synthesised by conventional DNA synthesis techniques using phosphoramidite monomers.

For imaging, one slide was inverted so that the chamber coverslip contacted the objective lens of an inverted microscope (Nikon TE200) via an immersion oil interface. A 60° fused silica dispersion prism was coupled optically to the back of the slide through a thin film of glycerol. Laser light was directed at the prism such that at the glass/sample interface it subtends an angle

of approximately 68° to the normal of the slide and subsequently undergoes Total Internal Reflection (TIR). The critical angle for glass/water interface is 66°.

Fluorescence from single molecules of DNA-Cy3, produced by excitation with the surface-specific evanescent wave following TIR, was collected by the objective lens of the microscope and imaged onto an Intensified Charge Coupled Device (ICCD) camera (Pentamax, Princeton Instruments, NJ). The image was recorded using a 532nm excitation (frequency-doubled solid-state Nd:YAG, Antares, Coherent) with a 580nm fluorescence (580DF30, Omega Optics, USA) filter for Cy3. Images were recorded with an exposure time of 500ms at the maximum gain of 10 on the ICCD. Laser powers incident at the prism were 50mW at 532nm.

Single molecules were identified by single points of fluorescence with average intensities greater than 3x that of the background. Fluorescence from a single molecule was confined to a few pixels, typically a 3x3 matrix at 100x magnification, and had a narrow Gaussian-like intensity profile. Single molecule fluorescence was also characterised by a one-step photobleaching process in the time course of the intensity and was used to distinguish single molecules from pixel regions containing two or more molecules, which exhibit multi-step processes.

To count molecules, a threshold for fluorescence intensities was first set to exclude background noise. For a control sample the background was essentially the thermal noise of the ICCD measured to be 76 counts with a standard deviation of only 6 counts. A threshold was arbitrarily chosen as a linear combination of the background, the average counts over an image and the standard deviation over an image. In general, the latter two quantities provide a measure of the number of pixels and range of intensities above background. This method gave rise to threshold levels which were at least 12 standard deviations above the background with a probability of less than 1 in 144 pixels contributing from noise. By defining a single molecule fluorescent point as being at least a 2x2 matrix of pixels and no larger than a 7x7, the probability of a single background pixel contributing to the counting was eliminated and clusters were ignored.

In this manner, the surface density of single molecules of DNA-Cy3 was measured at approximately 500 per 100 µm x 100 µm image or 5 x 10^6 cm^2.

CLAIMS

1. A device comprising an array of polynucleotide molecules immobilised on a solid surface, wherein each molecule comprises a polynucleotide duplex linked via a covalent bond to form a hairpin loop structure, one end of which comprises a target polynucleotide, and the array has a surface density which allows the target polynucleotides to be individually resolved.

2. A device according to claim 1, wherein immobilisation to the solid surface is via covalent attachment.

3. A device according to claim1 or claim 2, wherein the array has a density of at least 10^3 molecules/cm^2.

4. A device according to any preceding claim, wherein the array has a density of at least 10^6 molecules/cm^2.

5. A device according to any preceding claim, wherein the spacing between adjacent target polynucleotides on the array is at least 100nm.

6. A device according to any preceding claim, wherein the polynucleotide molecules are of DNA.

7. Use of a device according to any of claims 1 to 6, in an analysis procedure to determine the sequence of the target polynucleotide.

8. Use according to claim 7, wherein the procedure is genotyping.

9. Use according to claim 7 or claim 8, wherein the target polynucleotides are individually resolved by optical microscopy.

10. A method for the preparation of a device according to any of claims 1 to 6, comprising ligating a target polynucleotide to the 5' end of a first molecule capable of forming a duplex as defined in claim 1, and immobilising the first molecule to the solid surface either before or after ligation.

11. A method according to claim 10, wherein the immobilising is after the ligation of the target polynucleotide.

12. A method according to claim 10 or claim 11, wherein the target polynucleotide is in the form of double-stranded DNA, ligation is between one strand of the DNA and the first molecule, and the other strand is removed after ligation.

13. A method according to claim 12, wherein a further polynucleotide is hybridised to the first molecule with a one or more base gap between the further polynucleotide and the 3'-end of the first molecule, ligation is between the double-stranded DNA and the 5'-end of the first molecule and the further polynucleotide and hybridisation is subsequently disrupted to remove the further polynucleotide to form the target polynucleotide.

14. A method according to any of claims 10 to 13, wherein the 5'-end of the first duplex is phosphorylated and the target polynucleotide is dephosphorylated prior to ligation.

SEQUENCE LISTING

<110> Solexa Ltd.

<120> Polynucleotide Arrays and Their Use in Sequencing

<130> REP06208WO

<140> (not yet known)
<141> 2001-01-30

<150> 0002310.1
<151> 2000-02-01

<160> 2

<170> PatentIn Ver. 2.1

<210> 1
<211> 21
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence:
Oligonucleotide

<220>
<221> misc_structure
<222> (1)..(20)
<223> N = cytosine with a fluorescent Cy3 group
attached. M = thymine with hexaethylene glycol
attached.

<400> 1
nctgctgaag cgtcggcagg m 21

<210> 2
<211> 13
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence:
Oligonucleotide

<220>

<221> misc_structure
<222> (1)..(13)
<223> N = adenine with hexaethylene glycol attached.

<400> 2
ncctgccgac gct 13